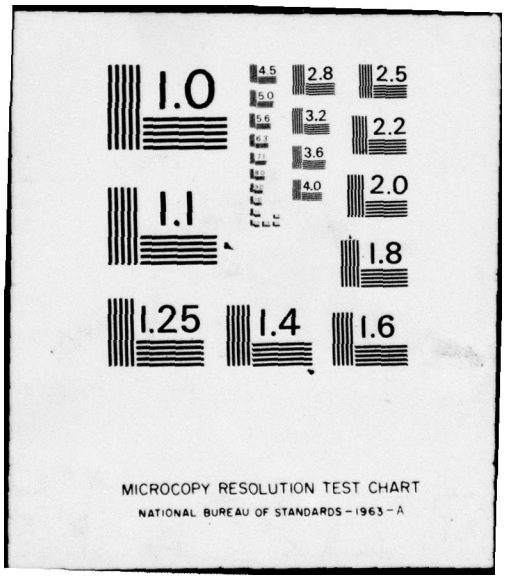
AD-A031 365 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/2
REFINED PREDICTION FOR LINEAR REGRESSION MODELS.(U)
AUG 76 J L HESS, R F GUNST AF-AFOSR-2871-75

UNCLASSIFIED AFOSR-TR-76-1125 NL

END

DATE FILMED
11-76


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

AFOSR - TR - 76 - 1125

REFINED PREDICTION FOR LINEAR REGRESSION MODELS

J.L. Hess and R.F. Gunst*
Department of Statistics
Southern Methodist University

ABSTRACT

Adequate prediction of a response variable using a multiple linear regression model is shown in this article to be related to the presence of multicollinearities among the predictor variables. If strong multicollinearities are present in the data, this information can be used to determine when prediction is likely to be accurate. A region of prediction, R, is proposed as a guide for prediction purposes. This region is related to a prediction interval when the matrix of predictor variables is of full column rank, but it can also be used when the sample is undersized. The Gorman-Toman (1966) ten variable data is used to illustrate the effectiveness of the region R.

1. INTRODUCTION

Prediction of future observations is one of the primary uses of an estimated linear regression model. Although a large number of papers and books have been written on the analysis of regression data, the emphasis in the literature is heavily weighted toward problems of model building and estimation of model parameters, and not on recommendations for using prediction equations. While these problems are all related, they do not necessarily place the same demands on the estimated model.

---

*This research was sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-75-2871.

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

Currently much of the statistical literature on linear regression is focussing on properties of biased regression estimators. Notable articles include James and Stein (1961), Hoerl and Kennard (1970), Marquardt (1970), Lindley and Smith (1972), Hawkins (1973), and Webster, Gunst, and Mason (1974). Biased estimation is receiving such prominence due to the realization that multicollinearity among the predictor variables (defined in Section 2) tends to severely distort the least squares estimates of the regression parameters. This in turn can result in poor prediction of future responses. Subset selection procedures likewise are not immune to distortion in the presence of multicollinear data.

Underlying this need for good parameter estimates is the assumption that the fitted model is to be used to predict over a wide region of interest of the predictor variables, perhaps an entire rectangular region defined by the extreme values observed on each predictor variable. This may be unduly stringent assumption as Hocking (1976) discusses from a variable selection viewpoint. In other words, frequently it is not necessary to predict over such a wide region. When this is so, accurate predictions may be possible despite uncertainties about the goodness of individual parameter estimates.

The purpose of this paper is to better identify when prediction is likely to be accurate with multicollinear data. Specifically, this paper was stimulated by three problems noticed by Owen and Reynolds (1968) in their development of a prediction equation for estimating engineering man-hours for proposed aircraft programs:

1) they decided to include "no more than 12" predictor variables from a total of about 60 possible ones since only 23 observations

on the response variable (engineering man-hours) were available;

2) a backward elimination (Draper and Smith (1966), Chapter 6) procedure was performed to further reduce the number of predictor variables to eliminate "the possibility of the accidental deletion of a significant variable due to its interaction with other variables"; and

3) the authors concluded that "some limits of extrapolation for formulas should be a primary objective of future studies."

We will address each of these problems in subsequent sections, not with the goal of providing final, definitive solutions to them, but rather to show how each affects the estimation of the regression parameters and the use of the resulting prediction equation. We do not intend to argue that any particular estimator is the best one to use with multicollinear data. We will, however, point out some advantages of using a principal component estimator to obtain a prediction equation.

2. LEAST SQUARES PREDICTION EQUATIONS

In this section we will examine problems 2) and 3) of Owen and Reynolds (1968) which were listed in the previous section. Suppose the assumed linear regression model is written as

$$\underline{Y} = \beta_0 \underline{1} + X\underline{\beta} + \underline{\varepsilon}, \tag{1}$$

where $\underline{Y}$ is an (n×1) vector of observations on the response (dependent) variable, $\underline{1}$ is an (n×1) vector of ones, $X = [\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_p]$ is an (n×p) full column rank matrix of predictor (independent) variables, β_0 is an unknown constant, $\underline{\beta}$ is a (p×1) vector of unknown regression parameters, and $\underline{\varepsilon}$ is an (n×1) vector of unobservable random error terms with

$\underline{\varepsilon} \sim N(\underline{0},\sigma^2 I)$. For simplicity, we assume that the elements of $\underline{X}_j$ are standardized so that $\underline{X}_j'\underline{1} = 0$ and $\underline{X}_j'\underline{X}_j = 1$ for $j = 1,2,\ldots,p$. Finally, model (1) is assumed to adequately represent the response variable although some of the predictor variables may not be needed for adequate prediction.

The latter two problems cited by Owen and Reynolds result from inadequacies in the data used to estimate the model parameters; in particular, from multicollinearities in the data. A multicollinearity can be defined as a linear combination of the columns of X that is nearly zero. This implies that X'X is nearly singular. A multicollinearity is not necessarily due to some variables being redundant in the specification of the model, but they may be redundant <u>for the data collected</u>.

Redundant model variables, those variables that will be redundant for all samples of data, can and should be deleted from the model since they serve only to inflate the variance of predicted responses (see, e.g., Hocking (1976)). If the redundancy is inherent only in the particular data sampled, it is dangerous to remove them from the predictor since the estimated model may then be biased when future responses are predicted. Yet multicollinearities tend to cause the deletion of one or more of the multicollinear variables merely because they are involved in multicollinearities, not because they are worthless predictor variables.

To see this latter point, denote the eigenvalues of X'X by $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$ and the corresponding eigenvectors by $\underline{V}_1,\underline{V}_2,\ldots,\underline{V}_p$. If there are one or more multicollinearities among the columns of X, one or more of the eigenvalues of X'X will be nearly zero. For eigenvalues that are near zero, multicollinearities can be identified by noting that

$$\underset{-j}{V}'X'X\underset{-j}{V} = \ell_j \approx 0 \Rightarrow \sum_{i=1}^{p} V_{ij}\underset{-j}{X} \approx \underset{-}{0}. \tag{2}$$

Equation (2) shows that the eigenvector $\underset{-j}{V}$ corresponding to a small eigenvalue ℓ_j provides the coefficients for the linear combination of the columns of X causing a multicollinearity. Naturally, the larger elements in $\underset{-j}{V}$ identify the predictor variables most strongly multicollinear. Mason, et al. (1975) contains a more complete discussion of multicollinearities and the problems associated with them.

The least squares estimator of $\underset{-}{\beta}$ for the model specified by (1) is

$$\hat{\underset{-}{\beta}} = (X'X)^{-1}X'\underset{-}{Y}.$$

The variances and covariances of the $\hat{\beta}_i$ can be found from

$$\text{Var}[\hat{\underset{-}{\beta}}] = (X'X)^{-1}\sigma^2 \approx \sum_{j=1}^{p} \ell_j^{-1}\underset{-j}{V}\underset{-j}{V}'\sigma^2. \tag{3}$$

From (3) we can see that small eigenvalues in X'X will result in large variances and covariances for estimated parameters of variables involved in multicollinearities (those with large V_{ij} values in (2)).

When attempting to reduce the number of variables in the prediction equation, the t statistic commonly used to test H_0: $\beta_j = 0$ is

$$t = \hat{\beta}_j/(c_{jj}\ \text{MSE})^{1/2}, \tag{4}$$

where c_{jj} is the jth diagonal element of $(X'X)^{-1}$ and MSE is the estimate of σ^2 computed from the full model (1). Since the c_{jj} values of variables involved in multicollinearities tend to be large due to the small ℓ_j in (3), the t statistics corresponding to these variables tend to be small. This accounts for the tendency for variables to be deleted by some computer programs because of their "interaction with other variables".

Contrary to Owen and Reynold's supposition, backward elimination also suffers from this problem. Backward elimination deletes the variable with the smallest t statistic at each stage. Since multicollinear variables tend to have small t statistics, at least one multicollinear predictor variable is likely to be deleted from the model. See Gunst, et al. (1976) for an illustration of this property.

The problem of eliminating important predictor variables, problem 2) of the previous section, is thus directly related to multicollinearities in the data. Multicollinearities in the data used to estimate $\underline{\beta}$ may affect prediction even if all the predictor variables are used in the prediction equation. Write the least squares prediction equation as

$$\hat{Y} = \hat{\beta}_0 + \underline{u}'\hat{\underline{\beta}}, \tag{5}$$

where $\hat{\beta}_0 = \bar{Y}$ and $\underline{u}$ is a vector of values of the p predictor values which are standardized as in (1).

Since $\bar{Y}$ generally estimates β_0 well, (5) will be an adequate predictor of the response if

$$\underline{u}'\hat{\underline{\beta}} \approx \underline{u}'\underline{\beta}$$

for all values of $\underline{u}$ in some region of interest. Now $\underline{u}'\hat{\underline{\beta}}$ is an unbiased estimator of $\underline{u}'\underline{\beta}$, with variance

$$\begin{aligned} \mathrm{Var}[\underline{u}'\hat{\underline{\beta}}] &= \sigma^2\, \underline{u}'(X'X)^{-1}\underline{u} \\ &= \sigma^2 \sum_{j=1}^{p} \ell_j^{-1} (\underline{u}'\underline{V}_j)^2. \end{aligned} \tag{6}$$

It can be seen from (6) that $\mathrm{Var}[\underline{u}'\hat{\underline{\beta}}]$ will be unacceptably large for many points $\underline{u}$ if one or more of the ℓ_j are sufficiently small, or some of the $\underline{u}'\underline{V}_j$ are large.

A commonly known but infrequently used means of estimating the precision of prediction is by forming a $100(1-\alpha)\%$ prediction interval for the point $\underline{u}$:

$$\hat{Y} - t_{\nu}(\alpha/2)\cdot s \leq Y \leq \hat{Y} + t_{\nu}(\alpha/2)\cdot s, \tag{7}$$

where $t_{\nu}(\alpha/2)$ is the upper $100(\alpha/2)\%$ critical point of the t distribution with $\nu = n-p-1$ degrees of freedom, and $s = [(1+n^{-1}+\underline{u}'(X'X)^{-1}\underline{u})\cdot MSE]^{1/2}$. The width of this prediction interval depends on $\sum_{j=1}^{p} \ell_j^{-1}(\underline{u}'\underline{V}_j)^2$, as in (6).

Both (6) and (7) essentially depend on how small $\underline{u}'\underline{V}_j$ is relative to ℓ_j. If ℓ_j is extremely small, then $\underline{u}'\underline{V}_j$ must also be small or prediction will be poor. These considerations suggest the definition of a "region of predictability" wherein prediction would be expected to be suitably accurate. One such region can be defined as

$$R = \{\underline{u}: \ |\underline{u}'\underline{V}_j| \leq c_j, \ j=1,\ldots,p, \text{ and } a_i \leq u_i \leq b_i\}, \tag{8}$$

where a_i and b_i are the minimum and maximum standardized values of the ith predictor variable, i.e., $a_i = \min\{X_{ki}; \ k=1,2,\ldots,n\}$ and $b_i = \max\{X_{ki}; \ k=1,2,\ldots,n\}$ for $i=1,2,\ldots,p$.

Two methods for choosing the c_j are

(i) $c_j = \ell_j^{1/2}$, or

(ii) $c_j = \max\{|\underline{w}_i'\underline{V}_j|, \ i=1,2,\ldots,n\}$, where $\underline{w}_i'$ is the ith row of X.

Method (i) insures that $\ell_j^{-1}(\underline{u}'\underline{V}_j)^2 \leq 1$, while (ii) bounds $\underline{u}'\underline{V}_j$ by the largest of the values for the points $\underline{w}_i'$ used to estimate $\underline{\beta}$. Each of these methods can be interpreted as requiring that the prediction equation only be used in regions for which data has been collected; i.e., the requirements (i) and (ii) limit extrapolation. If one wishes to predict outside R, the predicted values must be cautiously used, but this does indicate a partial response to point 3) of Owen and Reynolds.

By far the worst prediction will occur for points which have values of $|\underline{u}'\underline{V}_j|$ that are large for small values of ℓ_j. Suppose r multicollinearities have been detected by a careful examination of the ℓ_j and $\underline{V}_j$, as well as possibly other procedures such as investigating the "correlation" matrix, X'X, or the variance inflation factors (Marquardt (1970), Marquardt and Snee (1975)). The c_j, a_i, and b_i could then be relaxed in (8) for the first (p-r) directions $\underline{u}'\underline{V}_j$, j=1,2,...,p-r. In these directions extrapolation could be allowed with the knowledge that (7) would still provide reasonable bounds. These ideas will become even more important with the discussion of problem 1) in Section 3.

Note that R is based solely on sample information, information available to the data analyst at the time he wishes to make a prediction. If (8) is not satisfied, prediction may--and sometimes will--be accurate since (5) is an unbiased estimator of $\beta_0 + \underline{u}'\underline{\beta}$. Prediction for $\underline{u} \in R$ provides the assurance that the prediction equation is suitably precise.

If variable selection procedures are used to reduce the number of variables in the model, prediction will be adequate provided (2) holds for the points at which prediction is desired. This implies that $\underline{u}'\underline{V}_j \approx 0$ for these new points. But this restriction is again in the form of a region (8) with the c_j chosen suitably small for j = p-r+1,p-r+2,...,p. Thus if a region of predictability of the form (8) is constructed, least squares estimation and variable selection techniques will yield prediction equations which are accurate despite the multicollinearities in the data used to estimate the parameters. Outside this region the predictor cannot be expected to perform well due to large variances of the predictor or bias due to erroneously deleting important predictor variables.

3. A PRINCIPAL COMPONENT PREDICTOR FOR UNDERSIZED SAMPLES

Owen and Reynold's first problem, having to use only about 12 of 60 possible predictor variables in their initial models, results from fewer observations than predictor variables being available for the analysis. The full rank analysis of (1) using least squares requires that $n>p$, a requirement not satisfied by their data.

There is a wide range of model-building problems that could be addressed at this point concerning specification of model (1), but it is not within the scope of this paper to do so. We merely wish to raise the obvious questions regarding the deletion of many potentially valuable predictor variables (i) subjectively, (ii) on the basis of a partial analysis of the response and a subset of the predictor variables, or (iii) by using a stepwise procedure such as forward selection (see Mantel (1970) for some objections to this technique for full rank models). One acceptable means of deleting variables prior to an analysis of the complete (assumed correct) model (1) is if there are model redundancies.

Rather than demanding a full rank analysis, generalized inverse estimators offer another option. The generalized inverse solution is generally presented in a discussion of singular X matrices (as in designed experiments) for which $n>p$ (see, e.g., Rao (1965), Searle (1971), or Theil (1971)). While the existence of this estimator of $\underline{\beta}$ and its estimability characteristics are well-known, its potential use with undersized samples ($n\leq p$) has not been fully explored. An exception to this statement is in the economic literature of simultaneous equations systems (Fisher and Wadycki (1971), Khazzoom (1975), Swamy and Holmes (1971)).

The particular generalized inverse estimator we will examine in this section is referred to in the literature (e.g., Massy (1965), Marquardt (1970)) as a principal components estimator. If $n>p$ and X has rank $p-r$ (corresponding to $\ell_{p-r+1} = \ell_{p-r+2} = \ldots = \ell_p = 0$), the principal component estimator is defined to be

$$\tilde{\underline{\beta}} = (X'X)^{-}X'\underline{Y} = V_L L_L^{-1} V_L' X'\underline{Y} \qquad (9)$$

where the generalized inverse of X'X is $(X'X)^{-} = V_L L_L^{-1} V_L'$, $V_L = [\underline{V}_1, \underline{V}_2, \ldots, \underline{V}_{p-r}]$, and $L_L = \text{diag}(\ell_1, \ell_2, \ldots, \ell_{p-r})$. It is often demonstrated that (9) is the least squares estimator of $\underline{\beta}$ subject to the constraints $V_0'\underline{\beta} = \underline{0}$, where $V_0 = [\underline{V}_{p-r+1}, \underline{V}_{p-r+2}, \ldots, \underline{V}_p]$.

With undersized samples (i.e., $n \leq p$) there are typically s very small eigenvalues of X'X in addition to the r zero ones. We propose, therefore, a generalization of (9) for undersized samples which is of the same form but with $V_L = [\underline{V}_1, \underline{V}_2, \ldots, \underline{V}_{p-s-r}]$. Again, it can be shown that this is the least squares estimator of $\underline{\beta}$ subject to the constraints $V_0'\underline{\beta} = \underline{0}$ and $V_s'\underline{\beta} = \underline{0}$, where $V_s = [\underline{V}_{p-s-r+1}, \underline{V}_{p-s-r+2}, \ldots, \underline{V}_{p-r}]$.

Our rationale for using this estimator of $\underline{\beta}$ stems from a different justification than the parameter constraints given above. This justification stresses the use of the actual information provided by the matrix of predictor variables and, as we shall see, again yields guidelines for the use of the resulting predictor.

Let H be an $(n \times n)$ matrix of eigenvectors of XX', partitioned as $H = [H_L : H_s : H_0]$. H_L is $n \times (p-s-r)$ and contains the eigenvectors of XX' corresponding to the eigenvalues in $L_L = \text{diag}(\ell_1, \ldots, \ell_{p-s-r})$, H_s is $(n \times s)$ and contains the eigenvectors corresponding to the eigenvalues in $L_s =$

diag($\ell_{p-s-r+1}, \ell_{p-s-r+2}, \ell_{p-r}$), and H_0 contains the eigenvectors corresponding to the n-p+r zero eigenvalues. For undersized samples, r, the number of zero latent roots of X'X, is generally equal to p-(n-1), so that H_0 contains only one vector. Then we can write (e.g., Good (1969))

$$\begin{aligned} X &= H_0 \Phi V_0' + H_s L_s^{1/2} V_s' + H_L L_L^{1/2} V_L' \\ &= X_0 + X_s + X_L, \end{aligned} \tag{10}$$

where Φ is an $(n-p+r)\times r$ matrix of zeros and $X_0 = H_0 \Phi V_0'$, etc. Since $X_0 = \Phi$ and $X_s \tilde{\approx} \Phi$ (since L_s contains small eigenvalues), we see from (10) that $X \tilde{\approx} X_L$. This emphasizes the point that the entire space of predictor variables has not been sampled, only a subspace that is primarily spanned by the eigenvectors in V_L. Inserting X_L in place of X in (1) and obtaining the principal component solution to the normal equations yields (9) with V_L defined as above. This argument can also be used to justify the use of a principal component estimator for the full rank model if multicollinearities are present since, then, $X = X_s + X_L \tilde{\approx} X_L$.

The principal component prediction equation for undersized samples,

$$\tilde{Y} = \bar{Y} + \underline{u}'\tilde{\underline{\beta}}, \tag{11}$$

is biased. The bias of (11) can be written

$$\begin{aligned} B(\tilde{Y}) &= \underline{u}'\underline{\beta} - \underline{u}'(X_L'X_L)^{-}X_L'X\underline{\beta} \\ &= \underline{u}'\underline{\beta} - \underline{u}'V_L V_L'\underline{\beta}, \end{aligned} \tag{12}$$

and the variance of $\underline{u}'\tilde{\underline{\beta}}$ is

$$\begin{aligned} \mathrm{Var}[\underline{u}'\tilde{\underline{\beta}}] &= \underline{u}'(X_L'X_L)^{-}\underline{u}\sigma^2 \\ &= \sigma^2 \sum_{j=1}^{p-s-r} \ell_j^{-1} (\underline{u}'\underline{V}_j)^2. \end{aligned} \tag{13}$$

The variance term (13) does not suffer from having small eigenvalues as does (6), but (12) indicates that the predictor is generally biased. Note that if $V_0'\underline{u} = \underline{0}$ and $V_s'\underline{u} = \underline{0}$, $\underline{u}'\underline{\beta} = \underline{u}'V_L V_L'\underline{\beta}$, and (11) indeed turns out to be unbiased. This again reflects the fact that prediction should be accurate if we restrict the region of predictability to points in a general region that was actually sampled.

This discussion suggests a region similar to (8) within which prediction could be proposed, but outside of which prediction should not be recommended. Extrapolation can also be allowed in the space spanned by V_L. An evaluation of these recommendations is the subject of the next section.

4. AN ASSESSMENT OF R

In this section an example is presented to illustrate the potential benefits of using a region such as R as a guide in predicting. Again, a prediction interval of the form (7) is preferable to R when X is of full column rank and a sufficient number of observations are available to obtain a good estimate of σ^2. Otherwise, R can still be effectively used, as is now demonstrated.

The example concerns the ten variable data of Gorman and Toman (1966). A detailed analysis of this data, including a listing of the raw data, is given in Daniel and Wood (1971). Two analyses of this data are to be performed here: (i) a full rank analysis in which the first 15 of the n=36 data points are used to obtain a predictor, and (ii) an undersized sample analysis in which only the first 10 of the 36 data points are used. Each predictor is then used to predict the remaining observations.

With the full rank analysis, the smallest latent root of the standardized X'X matrix is $\ell_{10} = 0.0062$, with corresponding latent vector:

$$\underline{V}_{10}' = (.322, -.080, -.172, -.106, -.309$$
$$.787, .308, .050, -.029, .191).$$

From the discussion of Section 2, both (7) and (8) suggest that prediction should not be attempted unless $\underline{u}'\underline{V}_{10}$ is small. (For simplicity and ease of discussion, we are only considering one small latent root in this analysis. Since $\ell_9 = 0.019$, we may wish to consider the magnitude of $\underline{u}'\underline{V}_9$ as well). Using least squares, a predictor of the form (5) was constructed.

Figure 1 is a plot of $|Y_i - \hat{Y}_i|$ (labeled "RESIDUAL") versus $|\underline{u}'\underline{V}_{10}|$ (labeled "VS PRIME U") for the 36-15 = 21 data points not used to estimate the parameters in the prediction equation. The trend is clear: the magnitude of the residuals increases with the magnitude of $\underline{u}'\underline{V}_{10}$. While some moderate-sized residuals do occur with small magnitudes of $\underline{u}'\underline{V}_{10}$, there are no small residuals for large magnitudes of $\underline{u}'\underline{V}_{10}$.

Also evident from Figure 1 is the need to explore possible bounds on $\underline{u}'\underline{V}_{10}$. The two suggested in Section 2 turn out to be

(i) $|\underline{u}'\underline{V}_{10}| < \ell_{10}^{1/2} = 0.079$

and

(ii) $|\underline{u}'\underline{V}_{10}| < \max\{|\underline{w}_i'\underline{V}_{10}|\} = 0.047.$

While these bounds may be extremely effective for new values of $\underline{u}$ which satisfy (i) or (ii), the smallest value of $|\underline{u}'\underline{V}_{10}|$ for the 21 additional points is $|\underline{u}'\underline{V}_{10}| = 0.110$. Nevertheless, the trend in Figure 1 indicates that, at least qualitatively, a region such as R can be effective in assessing when prediction should <u>not</u> be attempted.

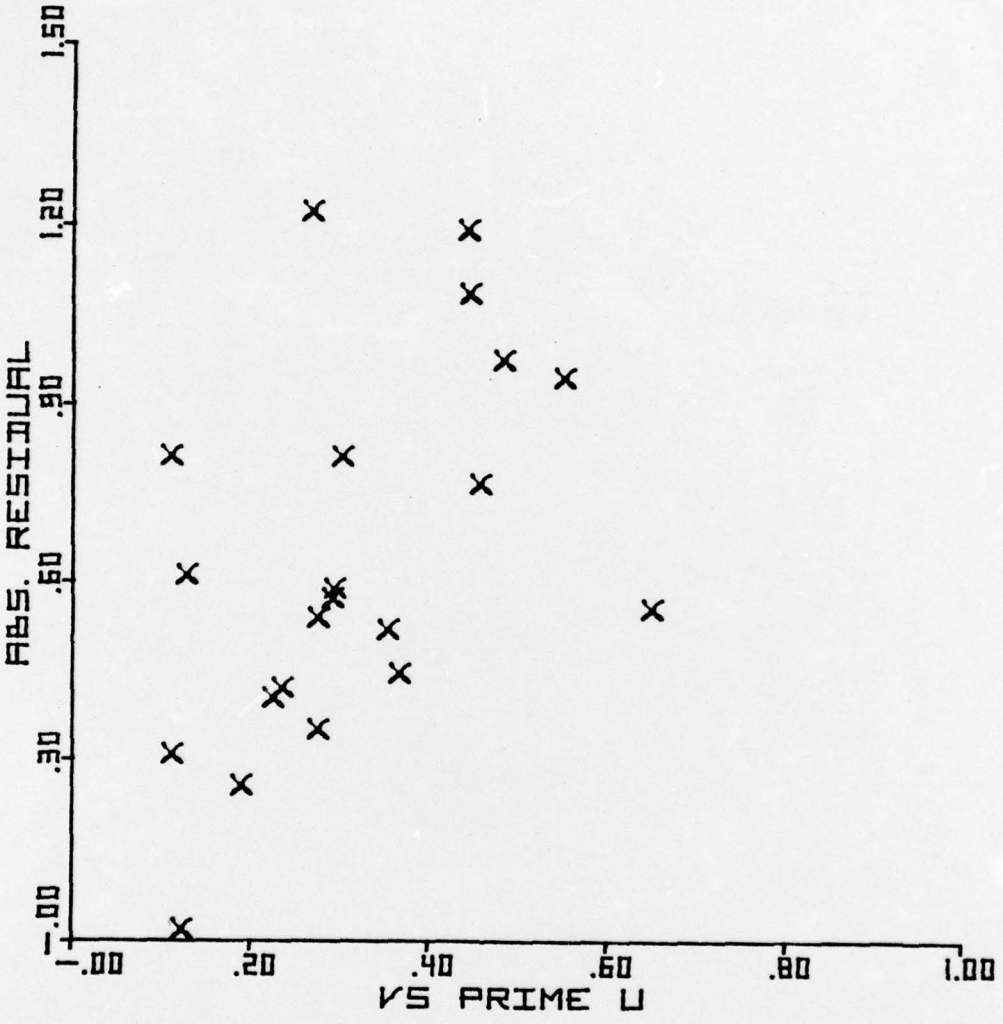


FIGURE 1. RESIDUALS OF GORMAN-TOMAN DATA BASED ON FULL RANK ANALYSIS.

For the undersized sample, $\ell_{10} = 0$ and $\ell_9 = 0.0104$. The corresponding latent vectors are

$$\underline{V}_9' = (.221, .046, -.131, .064, .562 \\ -.645, -.092, .148, .219, -.342),$$

and

$$\underline{V}_{10}' = (.661, -.085, -.367, -.059, .088 \\ .495, .290, .188, .160, -.139).$$

The latent vectors corresponding to the 8 remaining latent vectors were used to estimate $\underline{\beta}$ as in (9) and then form the prediction equation in (11). Figure 2 is a plot of the residuals, $|Y_i - \hat{Y}_i|$, of the remaining 36-10 = 26 data points (with a "+" indicating $|Y_i - \hat{Y}_i| \leq 0.75$, a "X" indicating $0.75 < |Y_i - \hat{Y}_i| \leq 1.50$, and " □ " indicating $1.50 < |Y_i - \hat{Y}_i|$) as a function of $|\underline{u}'\underline{V}_9|$ (labeled "VS PRIME U") and $|\underline{u}'\underline{V}_{10}|$ (labeled "VO PRIME U"). Again the trend is clear: smaller residuals occur predominantly with smaller values of <u>both</u> $|\underline{u}'\underline{V}_9|$ and $|\underline{u}'\underline{V}_{10}|$.

5. SUMMARY

The intent of this paper is to focus attention on an aspect of regression analysis that is often overlooked when the resulting prediction equation is employed. Regardless of the sample size used to obtain estimates of model parameters (and particularly when the sample size is small), estimation is highly inaccurate outside a region generally defined by (8). Yet regions of this form are always available to the data analyst and can be very valuable as guides in predicting. The Gorman-Toman data illustrates that both in the full rank situation and the undersized sample case, a region R formed by considering $\underline{u}'\underline{V}_j$ for latent vectors $\underline{V}_j$ corresponding to zero or small latent roots of X'X was effective in identifying when

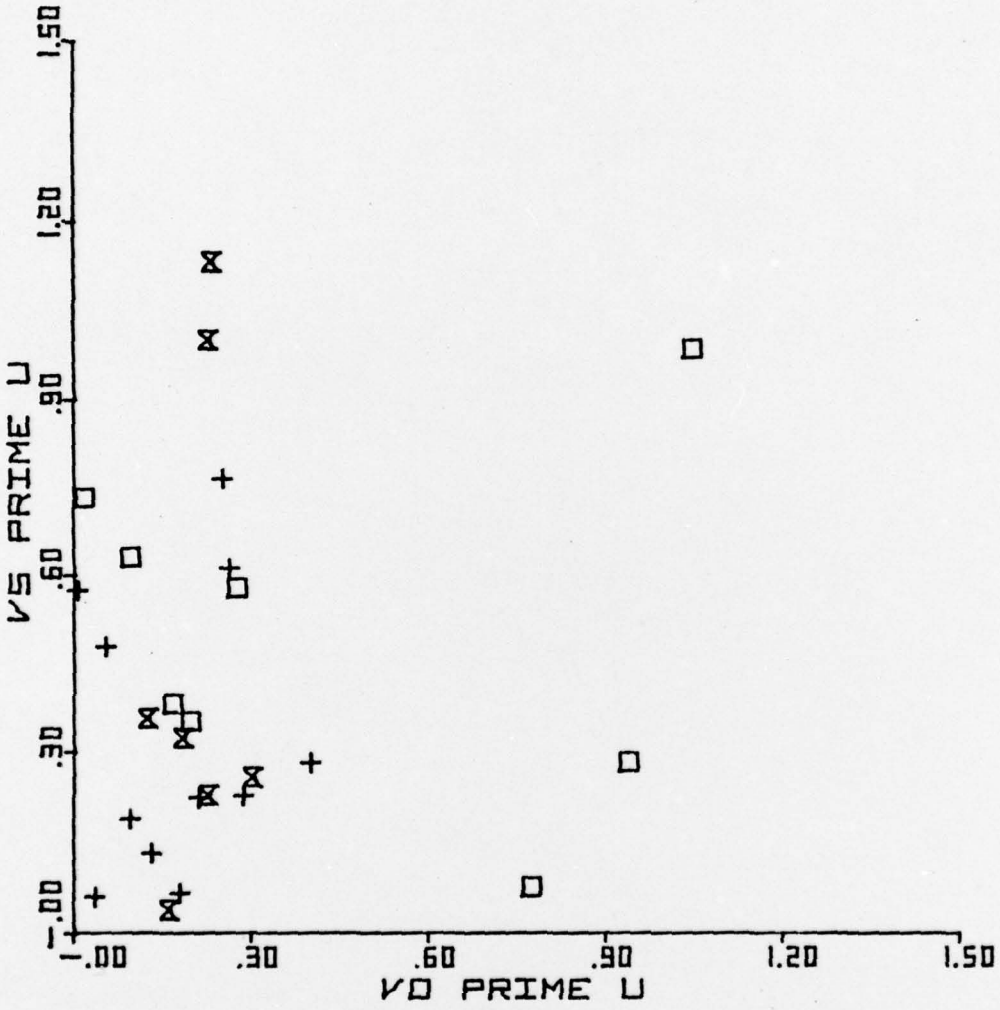


FIGURE 2. RESIDUALS OF GORMAN-TOMAN DATA BASED ON UNDERSIZED SAMPLE ANALYSIS.

prediction was likely to be inadequate. Further work in this area should concentrate on refining R; in particular, developing reasonable bounds, c_j, for (8) based on the information in X.

REFERENCES

[1] Daniel, C. and F.S. Wood, Fitting Equations to Data (Wiley Interscience, New York, 1971).

[2] Draper, N.R. and H. Smith, Applied Regression Analysis (John Wiley, New York, 1966).

[3] Fisher, W.D. and W.J. Wadycki, "Estimating a Structural Equation in a Large System," Econometrica, 39, 461-65 (1971).

[4] Good, I.J., "Some Applications of the Singular Decomposition of a Matrix," Technometrics, 11, 823-31 (1969).

[5] Gorman, J.W. and R.J. Toman, "Selection of Variables for Fitting Equations to Data", Technometrics, 8, 27-51 (1966).

[6] Gunst, R.F., J.T. Webster, and R.L. Mason, "A Comparison of Least Squares and Latent Root Regression Estimators," Technometrics, 18, 75-83 (1976).

[7] Hawkins, D.M., "On the Investigation of Alternative Regressions by Principal Component Analysis," Applied Statistics, 22, 275-86 (1973).

[8] Hocking, R.R., "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32, 1-49 (1976).

[9] Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Biased Estimation for Non-orthogonal Problems," Technometrics, 12, 55-67 (1970).

[10] James, W. and C. Stein, "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 361-79 (1961).

[11] Khazzoom, J.D., "An Exploratory Study of the Application of Generalized Inverse to ILS Estimation of Overidentified Equations in Linear Models," Tech. Report No. 219, Department of Statistics, Stanford University (1975).

[12] Lindley, D.V. and A.F.M. Smith, "Bayes Estimates for the Linear Model," Journal of the Royal Statistical Society, Series B., 34, 1-18 (1972).

[13] Mantel, N., "Why Stepdown Procedures in Variable Selection," Technometrics, 12, 591-612 (1970).

[14] Marquardt, D.W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," Technometrics, 12, 591-612 (1970).

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR - TR - 76 - 1125
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): REFINED PREDICTION FOR LINEAR REGRESSION MODELS.
5. TYPE OF REPORT & PERIOD COVERED: Interim rept.
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): J. L. Hess and R. F. Gunst
8. CONTRACT OR GRANT NUMBER(s): AF-AFOSR-2871-75

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Southern Methodist University
Department of Statistics
Dallas, Texas 75275

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F 9769-05

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research (NM)
Bldg. 410
Bolling AFB, Washington D.C., 20332

12. REPORT DATE: Aug 76
13. NUMBER OF PAGES: 19

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
21p.

15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
AF-9769
Approved for public release: distribution unlimited
9769-05

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Regression Analysis
Multicollinearity
Undersized Samples
Prediction Equations

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Adequate prediction of a response variable using a multiple linear regression model is shown in this article to be related to the presence of multicollinearities among the predictor variables. If strong multicollinearities are present in the data, this information can be used to determine when prediction is likely to be accurate. A region of prediction, R, is proposed as a guide for prediction purposes. This region is related to a prediction interval when the matrix of predictor variables is of full column rank, but it can also be used when the sample is undersized. The Gorman-Toman (1966) ten variable data is

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 404198

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.

used to illustrate the effectiveness of the region R.

[15] Marquardt, D.W. and R.D. Snee, "Ridge Regression in Practice," The American Statistician, 29, 3-20 (1975).

[16] Mason, R.L., R.F. Gunst, and J.T. Webster, "Regression Analysis and Problems of Multicollinearity," Communications in Statistics, 4, 277-92 (1975).

[17] Massy, W.F., "Principal Component Regression in Exploratory Statistical Research," Journal of the American Statistical Association, 60, 234-56 (1965).

[18] Owen, D.B. and D.F. Reynolds, "An Application of Statistical Techniques to Estimate Engineering Man-Hours on Major Aircraft Programs," Naval Research Logistics Quarterly, 15, 579-93 (1968).

[19] Rao, C.R., Linear Statistical Inference and Its Applications (John Wiley, New York, 1965).

[20] Searle, S.R., Linear Models (John Wiley, New York, 1971).

[21] Swamy, P.A.V.P. and J. Holmes, "The Use of Undersized Samples in the Estimation of Simultaneous Equations Systems," Econometrica, 39, 455-59 (1971).

[22] Theil, H., Principles of Econometrics (John Wiley, New York, 1971).

[23] Webster, J.T., R.F. Gunst, and R.L. Mason, "Latent Root Regression Analysis," Technometrics, 16, 513-22 (1974).